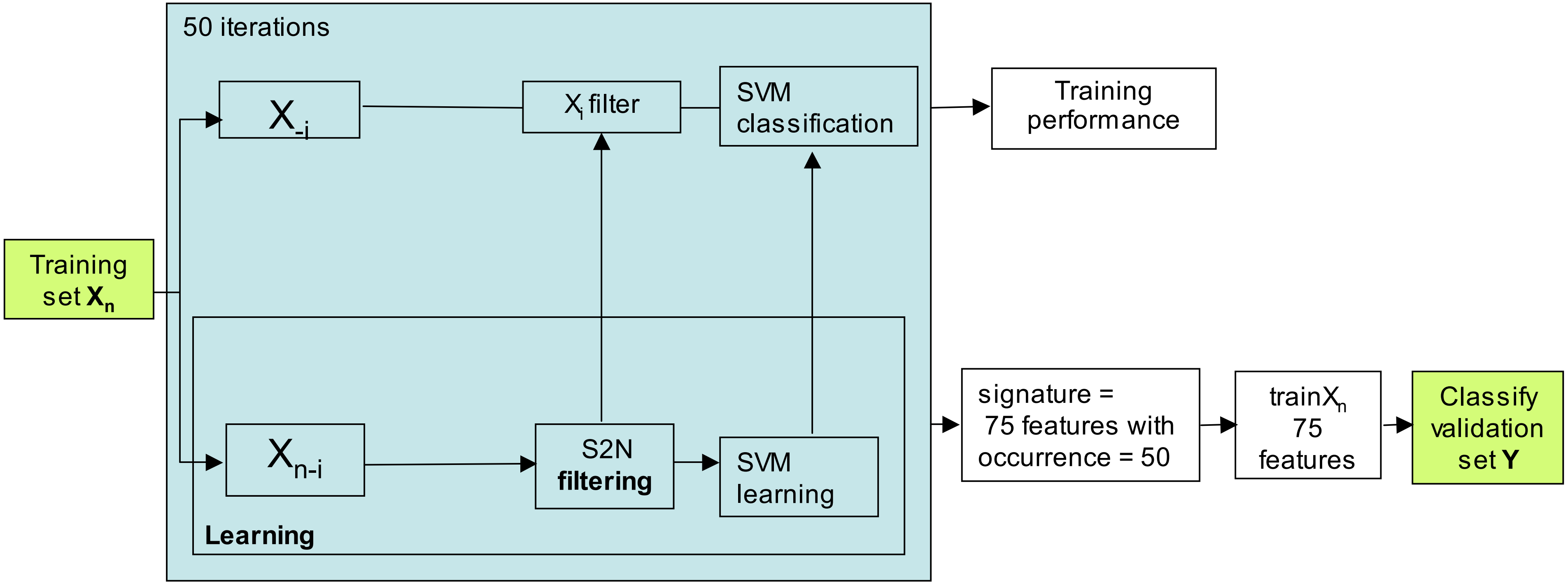
Bertucci and Orsetti et al.

Figure S1

Methodology used for supervised analysis



Supervised analysis

Identical data analyses were applied for array-CGH and gene expression profiles. Supervised analyses were done using the statistical computing environment R freely available at http://cran.R-project.org (R Development Core Team, 2007). We used a combination of signal-to-noise (S2N) ratio (Golub et al., 1999) to select informative features and support vector machine (SVM) for classification. The performance of the classifier was estimated by leave-one-out cross-validation (LOOCV). The S2N ratio is defined by $(\mu_0-\mu_1)/(\sigma_0+\sigma_1)$ where μ and σ represent the mean and the standard deviation of BAC or gene expression data for each class (here IDC vs. ILC). For feature (BAC or gene) selection observed ratios were compared to random ratios produced by 255 random permutations of the sample labels. Features were considered when the observed S2N exceed the random S2N at least 99% of the time ($p\leq0.01$) for BACs and 99.9% ($p\leq0.001$) for genes. Differential features were then tested with the classifier, which was trained on n-1 samples and this process repeated for each sample. We use the kernlab package (Karatzoglou et al., 2004) for SVM implementation. The applied SVM setting for BAC clone classification was polynomial kernel, C=1, $\gamma=1/75$, coef0=2^3, degree=5. For gene expression, we used the following parameters: polynomial kernel, C=2^{-5}, $\gamma=2^{-3}$, coef0=2^3, degree=5. Classifier performance was estimated by the accuracy (average of the n correct predictions). To avoid selection bias, the feature selection was implemented on the n-1 samples. Hence, informative features selected may vary according to each iteration. BACs and genes selected in all 50 iterations of LOOCV constitute the genomic and gene expression signatures.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. (1999). *Science*. Oct **15**;286(5439):531-7.

Karatzoglou A, Smola A, Hornik K. and Zeileis A.(2004). *Journal of Statistical Software.* **11**(9),1-20

Vapnik. (2000). *The Nature of Statistical Learning Theory*. Springer.